PROPOSITIONS POUR L'ENRICHISSEMENT SÉMANTIQUE DE CORPUS TEXTUELS

Coralie Reutenauer, Mick Grzesitchak, Evelyne Jacquey et Mathieu Valette
ATILF – CNRS, Nancy Université (UMR 7118)

ABSTRACT
The study implements a process of corpus annotation with semes relying on a textual semantics background. The incentive is to validate this process and also to analyze the additional information coming from this semantic annotation.

RÉSUMÉ
La présente étude met en œuvre une procédure d'annotation de corpus en traits sémantiques inspirée de principes de la sémantique textuelle. Elle cherche à évaluer d'une part la validité de l'annotation, d'autre part ses apports par rapport à une approche lexicale classique à partir d'un outil lexicométrique classique, le calcul des spécificités.

1 CONTEXTE ET OBJECTIFS
Le débat sur les métadonnées et l'enrichissement des corpus est soutenu. Tandis que la tradition française de la textométrie a longtemps considéré la forme comme unité de référence (cf. Brunet 2000, Mellet 2002 pour une discussion), le Traitement Automatique du Langage tend à lemmatiser systématiquement les corpus. Avec l'amélioration des techniques informatiques, on assiste à l'émergence de corpus multi-annotés et d'outils capables de traiter différents niveaux d'analyse, principalement morphosyntaxiques tels que les lemmes, les parties du discours et les catégories syntagmatiques (cf. par exemple le CorpusReader de Loiseau 2005). Si les outils d'annotation morphosyntaxique ont atteint une certaine maturité (Habert 2005), l'annotation sémantique reste peu dotée. Certes, le TAL et la Recherche d'Information confient parfois aux ontologies le soin de rendre compte de ce niveau, mais leur statut linguistique est très contesté (Slodzian 1999) – le peu d'attention qu'attirent ces ressources dans la communauté des statistiques textuelles et de la textométrie est sans doute l'indice de leur inadéquation. Récemment, une approche fondée sur l'exploitation de sèmes en guise de traits sémantiques a vu le jour. Un dictionnaire de sèmes qui se démarque de l'approche ontologique a été réalisé à partir d'une extraction depuis le *Trésor de la Langue Française informatisé* (Pierrel et Dendien 2003) (Valette *et al.* 2006, Grzesitchak *et al.* 2007, Valette 2008). Inspirés de la sémantique textuelle (Rastier 2001), les présupposés théoriques qui ont motivé la réalisation de cette ressource relèvent de conceptions partagées par la textométrie comme, par exemple, le primat accordé à la cooccurrence sur l'unité isolée (Mayaffre 2008).

L'objectif de cet article est d'évaluer, dans ce contexte, l'apport de l'annotation sémique pour la linguistique de corpus, en particulier pour la textométrie. L'expérience menée s'appuie sur une analyse lexicométrique classique et répandue, le calcul de spécificités. Elle repose sur

la confrontation d'un corpus de formes (sans annotation sémantique) au même corpus enrichi en traits sémantiques.

2 CORPUS : DU LEXICAL AU SÉMIQUE

2.1 Présentation du corpus

Le corpus utilisé est issu du discours journalistique. Il est constitué de 1587 articles de presse, tirés de deux quotidiens nationaux aux lignes éditoriales très contrastées, *Le Figaro* et *l'Humanité*. Les articles sélectionnés ont pour sujet la crise économique et financière ; ils couvrent la période de septembre 2008 à février 2009.

Le corpus se présente sous forme de deux versions parallèles : la version lexicale, d'un million d'occurrences de formes, et la version sémique (cf 2.2), de 23 millions d'occurrences de ce que nous qualifierons, en l'absence de validation systématique par le sémanticien, de "candidats-sèmes" par analogie aux *candidats-termes* de la terminologie. La taille du vocabulaire est du même ordre de grandeur dans les deux versions. Les informations principales sur la taille des deux versions du corpus sont récapitulées en figure 1.

	Total	*le Figaro*	*l'Humanité*
Nombre d'articles	1 587	928	659
Formes (version lexicale du corpus)			
Nombre d'occurrences	920 551	533 117	387 434
Nombre de formes	35 147	26 433	23 203
Candidats-sèmes (version sémique du corpus)			
Nombre d'occurrences	23 198 346	13 329 284	9 869 062
Nombre de candidats-sèmes	29 661	25 741	24 434

Figure 1 : Informations sur la taille du corpus

2.2 Constitution d'une version sémique du corpus

La constitution d'une version sémique du corpus est réalisée à partir d'une procédure mise au point par (Grzesitchak *et al*., 2007). Le schéma de la figure 2 récapitule les différentes étapes. Le corpus initial est étiqueté en morpho-syntaxe, lemmatisé et les mots-outils y sont éliminés. L'entrée correspondant à chaque lemme est recherchée dans une ressource lexicographique, le *Trésor de la Langue Française informatisé* (TLFi, Dendien & Pierrel, 2003). Seuls les substantifs, verbes, adjectifs et adverbes des définitions sont conservés. Chaque élément extrait de la définition est considéré comme un candidat-sème ; l'ensemble des candidats-sèmes issus d'une entrée constitue le sémème du lemme d'origine. Ce sémème est substitué au lemme en question dans le corpus. Ainsi, par substitution lemme par lemme, on obtient la version sémique du corpus.

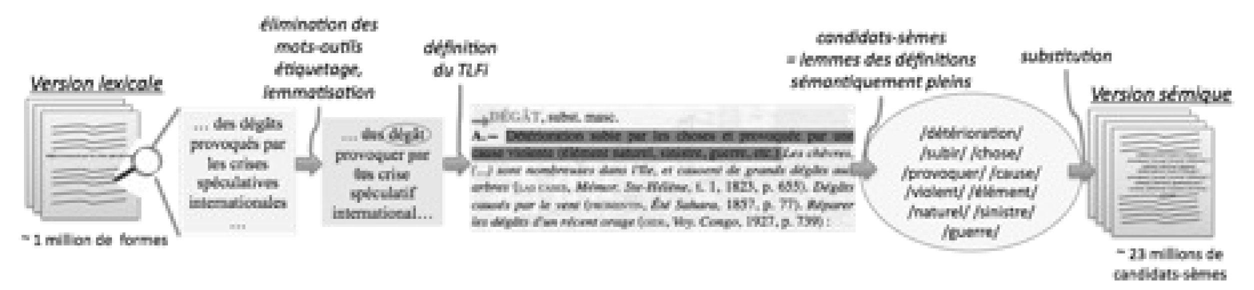


Figure 2 : Schéma de génération de la version sémique du corpus

3 ANGLES ET OUTILS D'APPROCHE DU CORPUS ANNOTÉ

Différents axes d'observation du corpus ont été retenus. Le développement de ces axes repose sur l'identification de contenu sémantique saillant, à l'aide du calcul des spécificités.

3.1 Outil mathématique : le calcul des spécificités

Le calcul des spécificités, décrit dans (Lafon, 1984), a pour but de déterminer le degré de surreprésentation ou de sous-représentation statistique d'une unité dans un sous-corpus par rapport à la totalité du corpus. Issu du modèle hypergéométrique, ce calcul utilise des comparaisons entre partie et tout. Pour une unité et un sous-corpus donnés, il nécessite les informations suivantes : le nombre d'occurrences de l'unité dans le sous-corpus ; le nombre d'occurrences de l'unité dans l'ensemble du corpus : la taille du sous-corpus ; la taille du corpus. Si l'unité est surreprésentée dans le sous-corpus (nombre d'occurrences dans le sous-corpus supérieur à la valeur modale), la valeur de la spécificité est calculée à partir de la probabilité d'avoir au moins le nombre d'occurrences observé ; cette spécificité est positive. Si l'unité est sous-représentée, la valeur est calculée à partir de la probabilité d'avoir au plus le nombre d'occurrences observé ; cette spécificité est négative. Les valeurs des spécificités sont des entiers.

Dans cette étude, le calcul des spécificités est implémenté par le logiciel de textométrie Lexico3 (Salem *et al.,* 2003). Les valeurs sont calculées au-delà d'un seuil de fréquence, fixé ici à 10. Toute unité (candidat-sème sur le plan sémique, forme sur le plan lexical) se voit affecter une spécificité si elle respecte les conditions de seuil.

3.2 Application au corpus

Le calcul des spécificités intervient dans deux approches, une approche globale et une approche locale. L'approche globale se situe à l'échelle d'un journal dans son ensemble. Elle vise l'étude de l'influence des lignes éditoriales de chaque journal. Elle est réalisée dans une perspective de validation de l'annotation sémique.

L'approche locale se focalise sur les paragraphes contenant un syntagme déterminé. Elle cherche à faire émerger des éléments issus du voisinage de l'unité ciblée et susceptibles de caractériser celle-ci. L'unité choisie ici est le syntagme *économie réelle*. Il est présent 176 fois dans 168 paragraphes. La taille du voisinage de cooccurrence est le paragraphe.

Les deux approches reposent d'une part sur une confrontation du plan sémique à une référence issue d'une évaluation intuitive, d'autre part sur une confrontation du plan sémique au plan lexical à l'aide des spécificités. Le calcul des spécificités est donc appliqué à la fois sur le plan sémique et sur le plan lexical.

Dans l'approche globale, le corpus est partitionné en deux selon la source, *L'Humanité* et *Le Figaro*. Ces deux sous-corpus servent tour à tour de référence pour le calcul de spécificité.

Notons que, par complémentarité des deux sous-corpus, une valeur positive sur une unité donnée dans un corpus correspond à la valeur opposée dans l'autre sous-corpus.

Dans l'approche locale, le sous-corpus de référence est, sur le plan lexical, l'ensemble des paragraphes contenant le syntagme *économie réelle,* et, sur le plan sémique, ce même ensemble de paragraphes converti en candidats-sèmes par la procédure d'annotation.

4 APPROCHE GLOBALE ET VALIDATION DE L'ANNOTATION SÉMIQUE

Les résultats obtenus dans l'approche globale, c'est-à-dire à l'échelle d'un journal, se présentent sous forme de listes de spécificités à la fois vastes et diversifiées, avec plus de 2000 formes sur le plan lexical et plus de 7000 candidats-sèmes pour un seuil de spécificité de 2. Deux approches manuelles ont été mises en place pour exploiter ces listes : l'observation des unités les plus spécifiques et un filtrage par catégories déterminées à la lecture.

4.1 Observation des unités les plus spécifiques

Le choix d'un seuil de spécificité élevé, de 20 sur le plan lexical et de 30 sur le plan sémique permet de réduire la liste considérée respectivement à quelques dizaines de formes lexicales et à une centaine de candidats-sèmes environ.. Les résultats présentés en figures (3a) et (3b) correspondent aux formes et candidats-sèmes les plus spécifiques de *l'Humanité.*

Forme	Spécificité	Forme	Spécificité	Forme	Spécificité
travail	≥50	syndicats	30	direction	22
CGT	≥50	communiste	29	des	22
salariés	≥50	sociales	28	Parti	21
salaires	≥50	social	27	capitalistes	21
PCF	48	on	27	pouvoir	21
capitalisme	43	emploi	26	public	21
gauche	41	politiques	25	dividendes	20
sociale	41	pôle	24	les	20
communistes	31	traité	23		
et	31	travailleurs	23		

Figure 3a : Formes lexicales les plus spécifiques de l'Humanité

Forme	Spf	Forme	Spf	Forme	Spf	Forme	Spf
organiser#v	≥50	engendrer#v	≥50	correctionnel#adj	40	syndiquer#v	34
propriété#subst	≥50	réorganisation#subst	≥50	échoir#v	40	formation#subst	34
prérogative#subst	≥50	syndic#subst	≥50	doctrine#subst	39	outil#subst	34
emploi#subst	≥50	communisme#subst	≥50	profit#subst	39	signer#v	33
employé#subst	≥50	rétribuer#v	≥50	utilisable#adj	39	gros#adj	33
commun#adj	≥50	indivision#subst	≥50	capitaliste#subst	39	égalité#subst	33
D=sociopolitique	≥50	critère#subst	≥50	concerner#v	39	catégoriel#adj	33
laborieux#adj	≥50	partageux#adj	≥50	condition#subst	39	entretenir#v	32
solidaire#adj	≥50	qualification#subst	50	gauche#adj	39	définir#subst	32
colonie#subst	≥50	voisinage#subst	50	inconfortable#adj	39	pension#subst	32
vénal#adj	≥50	communiste#adj	48	obligatoire#adj	38	peuplement#subst	32
employeur#subst	≥50	ouvrier#adj	48	venu#adj	38	dépendre#v	32
issir#v	≥50	communautaire#adj	47	militer#v	38	connexe#adj	32
culturel#adj	≥50	subvenir#v	47	paroisse#subst	37	D=droit	31
travailleur#subst	≥50	baser#v	46	voûter#v	37	besoin#subst	31
rémunération#subst	≥50	prôner#v	45	production#subst	37	tâcher#v	31
société#subst	≥50	pansage#subst	45	ligue#subst	37	entreprise#subst	31
dense#adj	≥50	ferrage#subst	45	recherche#subst	37	volontaire#adj	31
adaptation#subst	≥50	boeuf#subst	44	favoriser#v	37	profitable#adj	31
mondain#adj	≥50	expulsion#subst	43	affectation#subst	36	social#subst	31
strict#adj	≥50	but#subst	43	faculté#subst	36	régional#adj	31
mutation#subst	≥50	exercice#subst	42	bénéficier#v	35	protection#subst	30
amélioration#subst	≥50	patronal#adj	42	régulier#adj	35	notice#subst	30
salarier#v	≥50	besogne#subst	42	compromission#subst	35	honnête#adj	30
social#adj	≥50	remise#adj	41	subsistance#subst	35	notoire#adj	30
défavorisé#adj	≥50	utérus#subst	41	syndicat#subst	34	foetus#subst	30
communauté#subst	≥50	dilatation#subst	41	exercer#v	34	moyen#subst	30
classe#subst	≥50	col#subst	41	prisonnier#adj	34	richesse#subst	30
salarié#subst	≥50	capitalisme#subst	40	affirmé#adj	34		
marx#np	≥50	matériel#adj	40	sensé#adj	34		

Figure 3b : Candidats-sèmes les plus spécifiques de l'Humanité

Parmi les unités les plus spécifiques de *L'Humanité,* les orientations sociopolitiques du journal émergent nettement, pour les formes lexicales comme pour les candidats-sèmes : les problématiques des classes sociales, de la gauche, militantisme, syndicalisme et champ sémantique du travail et de l'emploi, sont très présentes. Par ailleurs, un certain nombre de candidats-sèmes renvoient à des notions moins classiques, plus latentes. C'est par exemple le cas de /prérogative#subst/ ou /vénal#adj/, de spécificité supérieure à 50, et dont l'équivalent est absent au niveau des formes les plus spécifiques. Cet enrichissement sur le plan sémique n'est néanmoins pas sans contrepartie : le bruit augmente au niveau des candidats-sèmes. Il provient de diverses sources :

57. de l'absence de filtrage domanial, à l'origine de candidats-sèmes non pertinents. Citons par exemple le cas d' /utérus#subst/, provenant de la définition de la forme lexicale *travail* rattachée au domaine de l'obstétrique (travail lors de l'accouchement). Ces traits non pertinents soulèvent la question du filtrage lors de l'annotation sémique, aussi bien domanial qu'interne à une définition.
58. de candidats-sèmes provenant du métalangage lexicographique, comme /concerner#v/
59. de candidats-sèmes non interprétables, par exemple en raison de leur caractère prédicatif (par exemple /favoriser#v/)

Les résultats sur les unités les plus spécifiques du *Figaro* indiquent également des contenus sémantiques en adéquation avec la ligne éditoriale du quotidien, avec un focus marqué sur les marchés et un regard tourné vers les puissances capitalistes. Le bruit est cependant plus important dans *Le Figaro* que dans *L'Humanité*.

Ainsi, l'étude des unités les plus spécifiques fait émerger des contenus sémantiques caractéristiques des deux journaux aussi bien sur le plan lexical que sémique, en conformité avec l'évaluation intuitive élaborée à partir de la lecture des articles. Cependant, la présence de bruit, accru lorsque la spécificité diminue, invite à raffiner l'approche de la liste de spécificités. La seconde approche, qui propose d'observer l'information lexicale et sémique au prisme de catégories définies manuellement, se situe dans cette optique.

4.2 Répartition en catégories

Lors de la constitution du corpus, le parcours des articles a permis de dégager des valeurs sémantiques caractéristiques de l'un ou l'autre des deux journaux. Ceux-ci ont servi à agencer les unités lexicales ou sémiques en catégories. Pour chaque catégorie, des formes et des candidats-sèmes pertinents sont sélectionnés. Cette sélection repose également sur un critère de rapprochement facile des formes et des candidats-sèmes. Par exemple, la sélection des formes lexicales *travailleur* et *travailleurs* fait pendant à celle de /travailleur#subst/, /travailleur#adj/ sur le plan sémique.

L'objectif est d'observer la convergence des formes et des candidats-sèmes sur des axes sémantiques majeurs. Notons que la sélection effectuée n'est pas exhaustive, il n'est donc pas question d'étudier l'expansion des unités du plan lexical vers le plan sémique, ni quantitativement, ni qualitativement.

Les résultats sont structurés en quatre grandes catégories : acteurs ; dimension nationale et internationale ; vocabulaire économique ; travail et activité. Ces grandes catégories sont subdivisées en sous-catégories. Une partie des résultats, extraite de la catégorie "acteurs", est présentée en figure (4). A chaque catégorie est associée une liste de candidats-sèmes ou de formes affectés de leur spécificité. Pour chaque journal, seules les spécificités positives sont indiquées. Par complémentarité, une spécificité positive pour un journal correspond à son opposé pour l'autre journal. Ainsi, la forme s*yndicat* est de spécificité +6 pour l'Humanité : elle sera donc de -6 pour le Figaro, et le coefficient 6 est reporté dans la colonne correspondant à *L'Humanité*.

Forme lexicale	Spécificité pour *Le Figaro*	Spécificité pour *L'Humanité*	Candidat-sèmes	Spécificité pour *Le Figaro*	Spécificité pour *L'Humanité*
Catégorie ACTEURS					
Syndicats					
syndicat		6	syndic#subst		≥50
syndicats		30	syndicat#subst		34
syndical		11	syndicalisme#subst		22
syndicale		12	syndical#adj		13
syndicales		12	intersyndical#adj		3
syndicalisme		4			
syndicaliste		12	militant#subst		4
syndicalistes		9	militer#v		38
syndicaux		6	militant#adj		25
intersyndicale		3			
Thibault		9	thibault#nam		9
délégué		10	délégué#subst		6
délégués		3			
CGT		≥50			
CFDT		9			
CFE-CGC		3			
CGC		4			
CFTC		6			
FO		6			
Partis					
PCF		48	parti#subst		11
PS		3			
UMP		5	ump#nam		6
Parti		21			
Acteurs socio-économiques et catégories socio-professionnelles					
agriculteurs		3	D=agriculture[1]		2
paysans		3	agriculteur#subst		6
ouvriers		5	ouvrier#adj		48
ouvrière		4	ouvrier#subst	12	
travailleur		4	travailleur#subst		≥50
travailleurs		23	travailleur#adj		8
salarié		7	salarié#subst		≥50
salariés		≥50	salarier#v		≥50
			salarié#adj		10
patron	7		patronat#subst		9
patrons		3			
patronat		10	patronal#adj		42

[1] D=... sert à désigner un domaine

Figure 4 : Spécificités des unités de la catégorie "acteur"

Les résultats obtenus indiquent d'une part une adéquation entre les observations humaines et les tendances indiquées par les spécificités, d'autre part une convergence entre plan sémique et lexical. Par exemple, la notion de syndicat apparaît comme très spécifique de *l'Humanité* aussi bien à travers les formes qu'à travers les candidats-sèmes. Ainsi, la convergence entre évaluation manuelle, plan lexical et plan sémique au niveau de grandes tendances valide l'annotation sémique. L'existence de différences plus fines au sein des catégories souligne un apport propre de l'annotation sémique, dont l'étude plus détaillée fait l'objet de l'approche locale.

5 APPROCHE LOCALE ET APPORTS DE L'ANNOTATION SÉMIQUE

Nous avons cherché à confronter le sens d'un mot-pôle, *économie réelle*, tel qu'il se dégage à la lecture à celui qui émerge d'une part à travers ses cooccurrents lexicaux et d'autre part à

partir d'un faisceau d'unités de sens issues du voisinage sémique. A la lecture des paragraphes, la crise économique apparaît comme une pathologie contagieuse ou comme une catastrophe naturelle se propageant de la sphère financière, considérée comme virtuelle, à la sphère industrielle, correspondant à l'économie dite réelle. Ces observations du lecteur ont servi par la suite à guider et à valider les analyses. Celles-ci portent dans un premier temps sur les unités les plus spécifiques du voisinage d'*économie réelle* et dans un second temps sur le voisinage filtré par des catégories déterminées à la lecture.

5.1 Unités les plus spécifiques du voisinage d'économie réelle

L'observation des listes de formes et candidats-sèmes les plus spécifiques du voisinage *d'économie réelle,* disponibles en figures (5a) et (5b), fait ressortir nettement une dimension économique et financière (présence par exemple des candidats-sèmes /budget/, /argent/, /capitaliste/, /économie/ sur le plan sémique, et des formes *financière, financier, profit* sur le plan lexical). De même, la sphère réelle apparaît à travers les unités les plus spécifiques, surtout sur le plan sémique, à travers des candidats-sèmes comme /chômage/, /bien/, /ressource/, /surproduction/. La notion de choc est également présente (/collision/, /répercussion/, /effondrement/ sur le plan sémique ; *impact* sur le plan lexical), de même que celle de propagation ou même de maladie (forme *contagion* ; candidats-sèmes /contagion/, /dysfonctionnement/ et /pathologique/). Les idées sensibles à la lecture se retrouvent ainsi au niveau des unités les plus spécifiques, sur le plan lexical et de façon encore plus marquée sur le plan sémique. Cependant, le nombre de formes ou candidats-sèmes associés à une idée donnée reste relativement limité, du fait de la taille volontairement réduite de la liste d'unités les plus spécifiques, d'où la mise en place d'une seconde approche des cooccurrents lexicaux et sémiques d'*économie réelle*. Cette seconde approche vise à établir des catégories partant d'idées dégagées de la lecture ou partant de l'observation des unités les plus spécifiques, à affecter des unités à ces catégories puis à confronter l'ensemble des représentants lexicaux et sémiques d'une même catégorie.

Forme (1/3)	**Spf**	**Forme (2/3)**	**Spf**	**Forme (3/3)**	**Spf**
l	20	financier	7	salaires	6
financière	15	conséquences	7	effets	6
impact	11	profits	7	revenus	6
crise	9	richesses	7	contagion	6
récession	9			dite	6

Figure 5a : Formes lexicales les plus spécifiques du voisinage d' "économie réelle" (seuil de spécificité de 6)

Candidat-sème	Spf	Candidat-sème	Spf	Candidat-sème	Spf
budget#subst	21	appréciable#adj	10	économique#adj	9
ressource#subst	16	capitaliste#adj	10	effondrement#subst	9
régir#v	15	collision#subst	10	enthousiasme#subst	9
argent#subst	14	contagion#subst	10	financier#adj	9
particulier#adv	14	décisif#adj	10	galaxie#subst	9
répercussion#subst	14	dysfonctionnement#subst	10	intense#adj	9
théâtre#subst	13	économie#subst	10	noeud#subst	9
bien#subst	12	époux#subst	10	pathologique#adj	9
chômage#subst	12	profond#subst	10	phénomène#subst	9
déterminant#adj	11	subit#adj	10	progressif#adj	9
diminution#subst	11	boursier#adj	9	retentissement#subst	9
néfaste#adj	11	craindre#v	9	rupture#subst	9
ralentissement#subst	11	D=dramaturgie	9	sous-production#subst	9
roi#subst	11	développement#subst	9	surproduction#subst	9

Figure 5b : Candidats-sèmes les plus spécifiques du voisinage d' "économie réelle" (seuil de spécificité de 9)

5.2 Filtrage par catégorie et émergence d'une forme sémantique

Les principales catégories choisies manuellement correspondent aux idées suivantes : la maladie ; le cataclysme ; le choc ou la brutalité ; la réalité ou, par opposition, la virtualité ; l'économie dans sa dimension matérielle. Les classes définies ont un degré de généralité variable. De plus, elles ne forment pas une partition : elles se superposent parfois et ne couvrent pas toutes les facettes sémantiques présentes dans l' "*économie réelle*". Certains candidats-sèmes sont donc affectés à plusieurs classes, tandis que d'autres ne rejoignent pas de classe particulière.

Pour constituer chaque catégorie, les listes de formes et de candidats-sèmes de spécificité supérieure à 2 ont été parcourues, avec un souci d'exhaustivité. L'affectation d'unités à certaines catégories s'est heurtée à des problèmes d'ambiguïté et à des cas d'incertitude. Des vérifications en contexte pour les formes et des recherches des formes génératrices pour les candidats-sèmes ont quelquefois été effectuées pour trancher sur l'affectation à une catégorie, mais cette procédure de contrôle n'a pu être systématisée, d'une part à cause d'usages variés des formes selon les contextes ou d'un trop grand nombre de formes génératrices, d'autre part en raison de la trop grande quantité de vérifications à faire.

A titre d'exemple, considérons les catégories suivantes : la catégorie 'maladie' (figure 6a) et la catégorie 'choc, brutalité' (figure 6b).

Trait	Spécif	Forme	Spécif
néfaste#adj	11	crise	9
contagion#subst	10	contagion	6
dysfonctionnement#subst	10	affectée	4
pathologique#adj	9	injectés	3
dépression#subst	8	affecter	3
trouble#subst	8	aggravée	3
crise#subst	7		
mal#subst	7		
physiologique#adj	6		
épidémie#subst	5		
infection#subst	5		
maladie#subst	5		
contagieux#adj	4		
remédier#v	4		
saignée#subst	4		
affecter#v	3		
bistouri#subst	3		
défaillir#v	3		
nuisible#adj	3		
psychose#subst	3		
soigner#v	3		

*Figure 6a : Catégorie **maladie** des unités spécifiques d' "économie réelle"*

Trait	Spécif	Forme	Spécif
effondrement#subst	9	choc	4
tarissement#subst	8	dommages	3
décrue#subst	8	éclate	3
dépression#subst	8	onde	3
choc#subst	7	fumée	3
violemment#adv	5	tempête	3
désagréger#v	4		
débris#subst	4		
déferler#v	3		
secousse#subst	3		
cataclysme#subst	3		
tempête#subst	3		
inondation#subst	3		

*Figure 6b : Catégorie **cataclysme** des unités spécifiques d' "économie réelle"*

Dans les deux cas, le nombre d'unités affectées à la catégorie est plus important sur le plan sémique que sur le plan lexical. De plus, certaines idées sensibles à la lecture mais sous-jacentes au niveau des formes lexicales apparaissent explicitement au niveau des candidats-sèmes. Par exemple, la maladie prend un caractère beaucoup plus prégnant et tangible avec

des candidats-sèmes tels que /pathologique/, /trouble/, /infection/, /épidémie/ ou encore /maladie/ ; de même, l'ébranlement et la violence liés à la crise, que seul *impact* reflète assez explicitement sur le plan lexical s'imposent avec force sur le plan sémique, avec des candidats-sèmes tels que /effondrement/, /heurt/, /brusque/, /violemment/ ou encore /secousse/. De façon générale, les catégories sont plus riches sur le plan sémique que sur le plan lexical, parce qu'elles contiennent plus de candidats-sèmes que de formes mais aussi, et surtout, parce que des idées perçues à la lecture sont exprimées clairement par les représentants sémiques alors qu'elles sont seulement sous-jacentes à travers les représentants lexicaux.

6 CONCLUSION

Cette étude décrit une procédure d'annotation en traits sémantiques de corpus, évaluée à travers la confrontation d'un corpus non annoté à son image annotée en candidats-sèmes. Les expériences réalisées indiquent une convergence entre l'évaluation intuitive de lecteur, le plan lexical et le plan sémique. Cette convergence se manifeste aussi bien à échelle globale (spécificités totales d'un journal par rapport à l'autre) que locale (focalisation sur le voisinage d'un mot-pôle). Les résultats valident ainsi la procédure d'annotation sémantique utilisée. Par ailleurs, l'approche en candidats-sèmes permet de faire émerger des formes sémantiques au voisinage d'un mot-pôle de façon plus marquée qu'au niveau lexical, d'une part en raison d'un accroissement des candidats-sèmes constitutifs de la forme sémantique, d'autre part en la profilant de façon plus fouillée que ne le fait le palier lexical de la forme présente.

L'enrichissement que propose l'annotation sémique est prometteur mais nécessite de se pencher sur le filtrage du bruit généré par l'annotation et sur le problème d'une polysémie inhérente à certains candidats-sèmes introduits. L'intégration d'informations domaniales ou encore la mise en place de représentations structurées des candidats-sèmes constituent des pistes susceptibles de réduire le problème. A travers ces développements, des perspectives plus larges s'ouvrent, comme la modélisation du sens pour la veille lexicale ou encore la détection de la néosémie.

7 RÉFÉRENCES

Brunet E. (2000). « Qui lemmatise dilemme attise ». *Scolia, 11e rencontres linguistiques en pays rhénan*, n°13, p. 7-32.

Dendien J. et Pierrel J.-M. (2003). « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence ». *TAL*, 44/2, p. 11-37.

Grzesitchak M., Jacquey E. et Valette M. (2007). « Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies ». *ARCo'07 – Cognition, Complexité, Collectif.* Acta-Cognitica, p. 227-235.

Habert B. (2005). « Portrait de linguiste(s) à l'instrument ». *Revue Texto ! Textes et cultures*, vol. X, n°4, disponible sur http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

Loiseau S. (2006). *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*. Thèse de Doctorat, Paris X-Nanterre.

Mayaffre D. (2008). « De l'occurrence à l'isotopie. Les cooccurrences en lexicométrie ». *Textes, documents numériques, corpus. Pour une science des textes instrumentée, Syntaxe & Sémantique*, 9, p. 53-74.

Mellet S. (2002). « Lemmatisation et encodage grammatical : un luxe inutile ? ». Lexicometrica, 3, 12.

Rastier F. (2001). *Arts et sciences du texte.* Paris : PUF.

Salem A., Lamalle C., Martinez W., Fleury S., Fracchiolla B., Kuncova A. et Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation.* Syled-CLA2T, Université de la Sorbonne nouvelle – Paris 3 http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW.

Slodzian M. (1999). « WordNet et EuroWordNet – Questions impertinentes sur leur pertinence linguistique ». *Sémiotiques,* n°17, p. 51-70.

Valette M., Estacio-Moreno A., Petitjean E. et Jacquey E. (2006). « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens ». *Verbum ex machina* (TALN 06), P. Mertens, C. Fairon, A. Dister, P. Watrin (éds). Cahiers du CENTAL, 2.1, UCL Presses Universitaires de Louvain. Volume 1. p. 357-366.

Valette M. (2008). « A quoi servent les lexiques sémantiques ? Discussion et proposition ». Dans M. Constant et al. (®ds.), Description linguistique pour le traitement automatique du français, *Cahiers du CENTAL*, n°5 – décembre 2008, PUL, p. 43-58.